Comparing Non-Verbal Vocalisations in Conversational Speech Corpora

Jürgen Trouvain[1] & Khiet P. Truong[2]

[1]Saarland University, Germany & [2]University of Twente, The Netherlands

E-mail: [1]trouvain [at] coli.uni-saarland.de & [2]k.p.truong [at] utwente.nl

Abstract

Conversations do not only consist of spoken words but they also consist of non-verbal vocalisations. Since there is no standard to define and to classify (possible) non-speech sounds the annotations for these vocalisations differ very much for various corpora of conversational speech. There seems to be agreement in the six inspected corpora that hesitation sounds and feedback vocalisations are considered as words (without a standard orthography). The most frequent non-verbal vocalisation are laughter on the one hand and, if considered a vocal sound, breathing noises on the other.

1. Introduction

Conversations do not only consist of spoken words but they also consist of non-verbal signals transmitted via the acoustic channel. Typical of these signals are that they often do not appear in dictionaries which is one of the reasons why people often have trouble writing down the signal's sound in orthographical form. Examples of these signals are laughter, coughs, breath sounds and feedback sounds such as "hmm-mm". We call these signals Non-Verbal Vocalisations (NVVs). Some of these vocalizations clearly have a communicative function and some are a result of the planning processes of speech production (what am I going to say next and how am I going to say it). As a consequence, NVVs are generally more present in spontaneous (conversational) speech than in carefully read aloud speech.

Research on NVVs in spontaneous conversational speech has been limited, which is partly due to the fact that NVVs are usually considered non-speech or 'garbage' sounds, especially from a technology point of view. Traditional automatic speech recognition (ASR) systems usually discard NVVs as non-speech sounds. However, researchers are becoming more aware of the importance of NVVs in spontaneous conversational speech and the need to model NVVs. Nowadays, ASR systems need to be able to recognize conversational speech and cope with NVVs. In addition, it is known that NVVs can carry communicative and affective meaning, which can be modelled for the development of spoken dialogue systems and emotion-aware systems.

Another possible reason for the limiting research performed on NVVs concerns the huge variability of NVVs. There is no clear definition of NVVs and there are no standard transcription and annotation protocols. These issues may have discouraged researchers to investigate NVVs in depth. Previous work on NVVs includes Ward (2006) in which a description of so-called conversational grunts in American English is presented. The focus of that study seems to cover only a part of the NVVs by our definition. We take on a broader view and include vocalizations such as laughter and audible breath sounds, which could play a role in dialogue. Our aim in this paper is to shed some light on the variability of NVVs.

The descriptive aims of study are to present various types of NVV and to sketch a scheme to structure various NVVs. The analytical part is to check which categories of NVV were considered in the different corpora and to find out i) differences in usage of NVV labels between different corpora and ii) frequencies of occurrence of various annotated NVV types. The results allow us to identify why and which NVVs can be important for communication research in conversational speech, and hence should be annotated with higher priority.

2. Types of non-verbal vocalisations

One problem of grouping and classifying NVVs is that the same or similar phonetic token can represent different NVVs. Breath intakes for example can be observed either as a vegetative sound or as part of a laugh or as a pragmatic signal with the meaning "I would like to have the turn." Here, we describe a number of possible types of NVVs.

2.1 Vegetative sounds

Vegetative sounds are not primarily communicative and not all are under voluntary control. Examples include snoring, moaning (e.g. in sports), swallowing sounds, chewing noises (with open or closed mouth), hiccup, coughing, sneezing, clearing the throat, yawning or panting (after physical exercise). Typically, vegetative sounds are not learned. However, there are vegetative sounds that require some level of learning such as spitting (e.g. cherry stones), lip smacking or producing an ingressive [s]. Probably the most frequent vegetative sound is audible inhalation. Audible exhalation sounds will also occur in conversation (not only after physical exercise).

Vegetative sounds can be used deliberately like clearing the throat ("ehem") to "say" that e.g. "I'm here now". Thus, deliberate vegetative sounds require pragmatic knowledge and the control of the vocal apparatus.

2.2 Affect sounds

Affect sounds include vocalisations such as laughing, weeping, cheering, crying loud and screaming. Conventionalised forms of these affect sounds include the deliberate use of moaning and yawning as well as imitations of coughing and snoring.

Schröder (2003) uses the more known term affect burst for affect sound, but then in a broader sense. It goes beyond the just described affect sounds incorporating also interjective words like "yippie" and "igitt".

2.3 Interjections as 'semi-words'

Sometimes the term interjection is used to indicate all kinds of NVVs with a paralinguistic character or at least those which "are tied to emotional or mental attitudes or states" (Wharton 2003). Sometimes interjections are meant to represent a certain word class, which would make them verbal vocalisations. Their debated linguistic status, the frequently unclear orthography and the fact that they often are not listed in dictionaries make them candidates for 'semi-words' (Wharton 2003).

Although there is no generally accepted definition of interjections they are often divided into primary and secondary interjections. The latter are words with an own meaning like "Damned! " or "Shit!" making them clearly verbal vocalisations. Primary interjections are e.g. "ouch" or "wow".

Onomatopoetic expressions like "miaow", "cuckoo", "knock-knock" can also be analysed as primary interjections, however, without any affective component. This is in contrast to interjections imitating environmental sounds in a less conventionalised way such as "woosh" or "bing". A further sub-category of primary interjections are affective words with an ungrammatical phonology such as "pst" or "shh" (no vowels) and "ts-ts-ts-ts" (clicks).

2.4 Feedback and filler sounds as 'semi-words'

Other 'semi-words' but without any affective component are hesitation sounds, also known as fillers or filled pauses such as "uh" or "uhm". Often they are regarded as disfluencies to which lengthened syllables (or syllable drawls) can be counted as well although this lengthening effect is not an independent vocalisation.

Another category of "semi-words" are sounds which function as feedback signals. They include humming signs like "hm" or "yeah" and "uhu". Usually they are used to backchannel but potentially also for asserting and other kinds of attitudinal expression.

2.5 Melodic utterances

A universal phonetic behaviour is the use of melodies with the own vocal apparatus. Melodies without text can be hummed, sung or whistled. We do not expect many of these utterances in conversation.

3. Distinctive dimensions

The same phonetic expression can be used for various functions. For instance breath sounds are primarily vegetative sounds. But breathing noises also play a role for laughter. Also an affect sound signalling startle usually involves a strong and sudden inhalation. Furthermore, audible inhalation can be used to signal to take the turn in a conversation. Another example is the humming sound (or neutral nasal consonant) which can be used for melodic purposes as well as for feedback signals and also for affective sounds signalling disgust but also pleasure – depending on its voice quality and its prosody. For this multi-functionality of NVVs we propose to describe them along four various distinctive dimensions of which one is binary ('vegetative') and three are not meant to be binary but continuous.

3.1 Vegetative dimension

Not all NVVs have a paralinguistic character and are uttered by the speaker to transport information. However, they contain extra-linguistic information about the speaker that can normally not be changed, e.g. coughs and sneezing can signal the status of the health or coughs can also be used for recognising the identity of a speaker. The vegetative dimension includes also not explicitly vegetative NVVs without any communication partner, e.g. affect sounds expressing pain.

3.2 Spelling dimension

There is no clear-cut border between NVVs in a narrow sense and semi-words. The decisive dimension to consider a vocalization as belonging to one of the semi-word classes or not seems to be the spelling dimension. Several times a continuum has been proposed reaching from 'raw' affect bursts (cf. Schröder 2003) or 'natural sounds' (cf. Wharton 2003) at the one end and secondary interjections at the other. At the one extreme reliable spelling of the expressed sounds is (nearly) impossible, on the other extreme the orthographic standard is rather clear. The spelling dimension also reflects the fact that NVVs at the non-spelling end are phonetically encoded by glottal rather than supra--glottal activities.

3.3 Affective dimension

Affect sounds and (most) interjections are defined by the affective dimension thus transporting a lot of information about the speaker and her/his attitudes and feelings in a very short time. Affective information is usually not present with vegetative sounds and filler sounds. Feedback sounds, however, can sometimes transport affective information.

3.4 Pragmatic dimension

Some NVVs act as pragmatic particles with functions for the management of the conversation. For instance feedback sounds such as 'backchannels' are indispensable for keeping a conversation fluent. Filler sounds can signal some problems with the self-management of the talker but it can also show upcoming new information. But also laughter and other affect sounds can be used as a feedback signal.

A summarization of the types of NVV as described in section 2 and the proposed dimensions in this section can be found in Table 1. It must be noted that the classification presented in types and dimensions is just a *sketch* for further theoretical considerations as well as empirical analyses.

Table 1: Gray areas and plus-signs indicate the presence and the intensity of the three continuous dimensions for the various types of NVVs (the binary dimension 'vegetative').

dimensions / types	veg.	spelling	affective	pragm.
vegetative sounds		-	-	-/+
deliberate veget. s.		+	+	+
affect sounds		-	+++	-
deliberate affect s.		+	++	-
imitative sounds		+	+++	-
melodic utter.		-	+++	-
interjections		++	++	-
fillers		++	-	+++
feedback sounds		++	-/+	+++

4. Differences in usage

Six different corpora of conversational English were inspected: ICSI meeting corpus (Janin et al. 2003), AMI (Carletta 2007), Switchboard (Godfrey & Holliman 1997), Diapix Lucid corpus (Baker & Hazan 2011), HCRC Map Task corpus (Anderson et al. 1991) and the Buckeye corpus (Pitt et al. 2007).

Annotations of the above mentioned NVV widely differ among corpora of conversational speech. All corpora consider the "semi-words" listed in sub-section 2.3 as words, although the orthography differs very much. It must be noted that a comparison is very hard due to different treatments of NVV annotations as tokens in the various annotation schemes but also due to various annotators, differences in conversational tasks and differences in microphones.

Laughter is always annotated in the corpora under inspection. However, speech-laughs were not always annotated as such (see table 2). Despite the various differences of the inspected corpora it seems obvious that annotated "laughs" is the predominant type of NVV in all corpora (cp. Fig. 1): more than 60% of all annotated NVVs in AMI and more than 40% in ICSI and Switchboard. However, the remaining three corpora show a remarkably low number of laughs, which can be attributed to a smaller amount of recorded data, the dyadic of multiparty character, and the conversational task.

The differences regarding breathing sounds are rather dramatic (see fig. 1). In the Buckeye corpus breath sounds are not a category at all whereas in AMI the transcription guidelines provide an appropriate annotation tag but it was extremely rarely selected (0.2% of all NVVs).

We understand that breath sounds in Switchboard were treated as the 'other' category which was named 'noise'.

Table 2: Table of occurrences of NVVs in various corpora. ‘N/A’ means that the vocalization was not explicitly mentioned in the transcription guidelines and was hence not considered by the transcribers. A zero ‘0’ means that the vocalization was mentioned in the transcription guidelines (and thus considered by the transcribers) but we cannot count these because there were not any or they were included in an explicit ‘Other’ category. ‘The rest’ means all the other annotated NVVs that did not fit one of our categories under inspection.

	Multiparty				Dyad							
	ICSI		AMI		Switchboard		Diapix		HCRC		Buckeye[1]	
N conversations	75		171		2438		57		128		255	
Duration	72h		100h		518h		7.3h		14.5h		37.8h	
	Abs	%	Abs	%	Abs	%	Abs	%	Abs	%	Abs	%
Laugh	12643	40.8	16477	61.0	22209	37.4	582	8.9	1002	5.3	1899	7.2
Speech-laugh	1017[2]	3.3	n/a	n/a	13503	22.7	333	5.1	n/a	n/a	1020	3.9
Breath	12465	40.2	57	0.2	0	0	3539	54.2	12280	64.8	n/a	n/a
Cough	256	0.8	1114	4.1	0	0	n/a	n/a	320	1.7	0	0
Clearing the throat	906	2.9	0	0	0	0	0	0	n/a	n/a	0	0
Lip smacking	n/a	n/a	3	0.0	n/a	n/a	1182	18.1	4512	23.8	n/a	n/a
Eating	39	0.1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Yawn	62	0.2	10	0	0	0	n/a	n/a	n/a	n/a	n/a	n/a
Sigh	22	0.1	47	0.2	0	0	0	0	n/a	n/a	0	0
Humming/Singing/ Whistling	47	0.2	85	0	0	0	n/a	n/a	n/a	n/a	n/a	n/a
Other	n/a	n/a	8888	33.0	23682	39.9	893	13.7	n/a	n/a	22661	86.3
The rest	3554	11.5	310	1.1	0	0	0	0	823	4.3	685	2.6
Total	31011	100	26991	100	59394	100	6529	100	18937	100	26265	100

[1] Only one person of the dyad was recorded and annotated

[2] Counts of segments (instead of separate words) spoken while laughing

Diapix and HCRC show the expected high number of breath sounds whereas ICSI shows a medium-scaled number. This rather disparate picture is also reflected in the plethora of the often detailed tags such as "inbreath", "outbreath", "long loud outbreath", "loud inhale", "strong exhale" etc.

When looking at the token frequency of selected NVV types it can be easily observed that laughter and breathing sounds dominate. Other NVVs like cough, clearing the throat, yawning etc (see table 2) show a rather low frequency of occurrence (with the exception of lip smacking for the HCRC map task corpus).

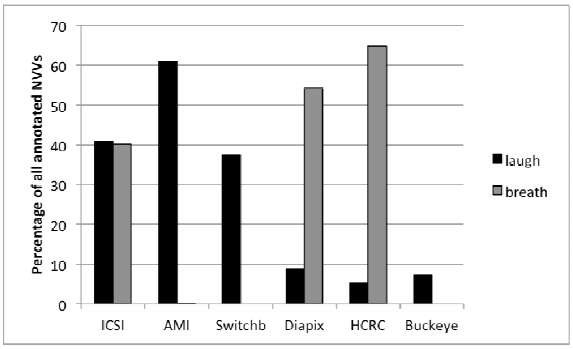

Fig1: Laughter and breathing sounds were the two main NVVs annotated in the inspected corpora. The graph shows the numbers of annotated "laugh" and "breath" relative to the total number of NVV for each corpus.

5. Concluding remarks

Our analysis of six corpora with conversational speech revealed that there is a *huge* disparity among the inspected corpora with regard to the annotation of NVVs. There seems to be agreement that 'semi-words' like feedback and filler sounds as well as interjections with a possible spelling should not be regarded as 'non-verbal' or 'non-speech'. It also turned out that laughter represents the category of NVV with the highest frequency of occurrence. However, there is disagreement about the status, the amount and the specification of breathing sounds. Other types of NVV such as coughing, eating sounds, yawning or melodic utterances either play only a minor role or are not yet explored in the inspected corpora of conversational speech. Although one could say that these sounds do not seem to be much dialogue-related, we do not recommend exclusion of these, as some of these sounds *can* be useful for dialog research. For example, a cough can contain speaker identity information and yawning or singing can be signals of tiredness or good mood.

Usually the details of the annotation of NVV depend on the goal of investigator's research. However, corpora of conversational speech provided for general research on how spoken interaction unfolds would also need a more detailed annotation of NVVs. Based on our investigations and with respect to future research we consider it worthwhile to have more consistent and detailed NVV annotations. In particular, research on turn-taking could benefit from consistent annotation of breath sounds which can also serve as additional signals for prosodic breaks in general.

The difficulty providing practically useful and theoretically valid definitions of NVV reflects the lack of knowledge about the acoustics as well as about the functions NVVs can serve. Some NVVs show similar phonetic shapes but serve different functions. For example, a schwa-sound or a neutral nasal consonant can occur as a token of each NVV type. It just depends on the glottal and sub-glottal activity (voicing, voice quality, intonation, respiration) *and* the context (syntactic position and articulatory isolation) that makes this sound have a certain interpretation. An analysis of additional annotations such as dialogue act annotations in which pragmatic functions like feedback, filler etc. are annotated could be helpful.

In order to provide a better basis for comparing different corpora a re-annotation of the NVV would be advisable. This would require a theoretical framework to put NVVs into a larger context of which here only a few points were discussed. A theoretical fundament backed with empirical data would also allow comparisons of NVVs between taken from experimental lab studies and spontaneous conversations.

Acknowledgements

This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet) and the UT Aspasia Fund.

6. References

Anderson, A.H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weintert, R. (1991). The HCRC Map Task Corpus. *Language and Speech* 34(4), pp. 351-366.

Baker, R, Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3), pp. 761-770.

Carletta, J.C. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation* 41(2), pp. 181-190.

Godfrey, J.J., Holliman, E. (1997). *Switchboard-1 Release* . Linguistic Data Consortium, Philadelphia.

Janin, A., Baron, D., Edwards, D., Ellis, D., Gelbart, D., Morgan, N. (2003). The ICSI meeting corpus. *Proceedings of ICASSP*, pp. 364-367.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release) [www.buckeye corpus.osu.edu] Columbus, OH: Ohio State University. (Distributor).

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication* 40 (1-2), pp. 99-116.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics and Cognition*, 14(1), pp. 129-182.

Wharton, T. (2003). Interjections, language and the 'showing'-'saying' continuum. *Pragmatics and Cognition*, 11(1), pp. 39-91.